《Python 程序设计与 R 语言》

1. 大作业题目

二手房数据分析

2. 作业内容

（1） 获取数据集：使用 Python 爬虫从贝壳或链家网站中获取北京市东城区、海淀区、通州区、怀柔区的在售二手房数据，并选择合适的方式进行存储数据。下面的数据分析均针对东城区、海淀区、通州区、怀柔区这四个区的数据进行。（20 分）

要求：**每个城区的数据不少于 1000 条，字段属性不少于 10 个**，提交的作业中包含：

1） 爬取方法作为报告的一个独立章节。

2） **爬虫代码和爬取到的原始数据文件**。

代码：指明在提交代码目录中所属文件名、模块名、函数名。（见：各功能源代码说明表示例）

原始数据文件：格式自选，可以是 SQL 数据文件或 csv 数据文件或其他格式的数据文件。

（2） 数据预处理：对采集到的数据集进行重复值处理、缺失值处理、异常值的检测与处理。（10 分）

要求：报告中要写明做了哪些预处理操作和预处理的步骤，**至少要进行重复值处理和缺失值处理**，提交的作业中应包含：

1） **预处理方法作为报告的一个独立章节；**

2） **处理代码**和**预处理后的数据文件**。

代码：指明在提交代码目录中所属文件名、模块名、函数名。（见：各功能源代码说明表示例）

预处理后的数据文件：格式自选，可以与原始数据文件不一致。

（3） 这四个城区在售的二手房有哪些特点，不限于从总价、单价、户型类别、楼层、建筑面积、建筑年代、电梯安装情况等属性进行统计性分析和可视化分析。例如，进行统计学分析时，可以使用箱线图分析并展示平均值、最大值、最小值等统计值的关系。（30 分）

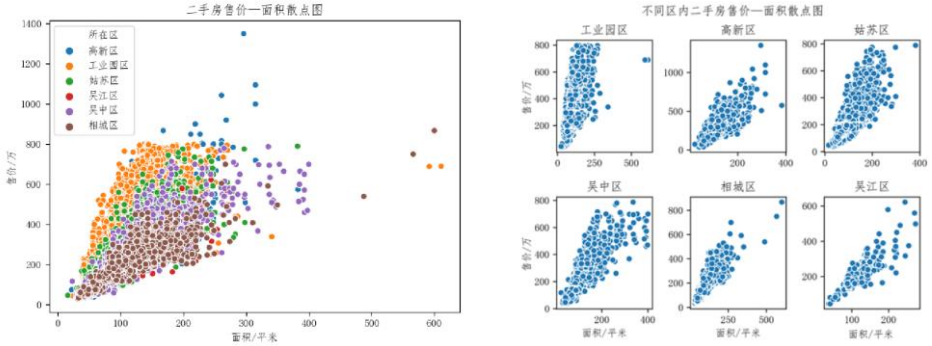
要求：**分析的数据特征不少于 5 个**，分析每种数据特征时**可视化图形种类不少于一种**，比如，分析二手房单价（5 个特征之一）时，如果使用箱线图，则需要分别绘制四个城区的箱线图。提交的报告中应包含：

1） 统计分析和可视化分析的思路和结论作为报告的一个独立章节，本章节中插入可视化的图形（给出图题）。

2）图形绘制代码。**（**指明在提交代码目录中所属文件名、模块名、函数名。见：各功能源代码说明表示例）

(4) 分析影响房价的主要因素有哪些，即分析房价与其他变量之间的关系。

例如：分析住房面积对二手房总价的影响，绘制住房面积与总价的散点图，对比分析四个城区之间的不同，并给出合理的分析。（20 分）

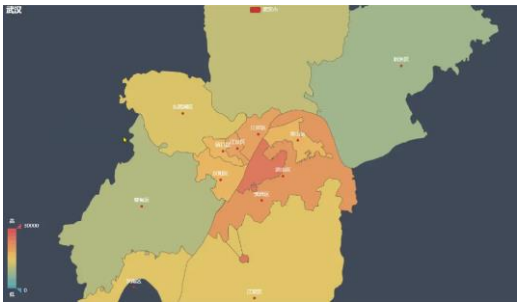


要求：**分析的变量不少于 1 种，可视化图形种类不少于两种。**

1） 分析的思路和结论作为报告的一个独立章节，本章节中插入可视化的图形（给出图题）。

2） 图形绘制代码。指明在提交代码目录中所属文件名、模块名、函数名。见：各功能源代码说明表示例）。

(5) 在地图上绘制这四个城区中在售二手房单位面积价格平均值的热力图和在售二手房总金额的热力图，分析二手房价格的地理分布特征。例如，下图中越红的区域表明单位房价越高。（20 分）



要求：报告中应有上述要求的**两个热力图，**并给出适当的分析。

1） 分析的结论作为报告的一个独立章节，章节中插入可视化的图形（给出图题）。

2） 图形绘制代码。**（**指明在提交代码目录中所属文件名、模块名、函数名。见：各

功能源代码说明表示例)。

(6) 选做题：使用任意一种算法对数据进行分析建模，预测房价走势。例如，使用机器学习中的线性回归模型，将数据划分训练集和测试集，选择一个或多个相关的数据特征对模型进行训练，需要解释你选择的特征为什么与房价相关，最后使用测试集评估模型的准确率。(10 分)

要求：

1) **预测的思路和结论作为报告的一个独立章节，给出体现预测结果的数据（图、表均可）**

2) 预测的代码。**(**指明在提交代码目录中所属文件名、模块名、函数名。见：各功能源代码说明表示例)。

3. 整体要求

(1) 应使用 Python 或 R 语言实现上述功能（也可以 python 与 R 混用），代码中给出必要和合理注释。

(2) 报告要求不少于 3000 字，避免大篇幅的展示程序代码，可以展示部分关键代码（代码不计字数)。

(3) 提交作业时应将源代码、数据文件和报告放至压缩包，一并提交。

报告内容：至少包括上述每一步要求的部分。

数据文件：包括：原始数据文件和预处理后的数据文件。

源代码：全部源放在一个目录下，并给出 readme，说明如何执行可以得到上述每个要求的功能。

各功能源代码说明表

功能	文件名	模块名	函数名
获取数据集	a.py	crawler	（全部）
数据预处理	b.r	---(无)	X():去除重复 Y()：处理缺失
统计性分析及可视化	c.py		
影响房价因素分析	d.r		
热力图分析	e.py		
房价预测	f.py		